AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 111:283–289 (2000)

Testing Type II Error Rates in Biological Anthropology

STEVEN N. BYERS*
Department of Anthropology, University of New Mexico, Albuquerque, New Mexico 87131

KEY WORDS Type II errors; hypothesis formulation; statistical tests

ABSTRACT This paper presents a look at the underused procedure of testing for Type II errors when "negative" results are encountered during research. It recommends setting a statistical alternative hypothesis based on anthropologically derived information and calculating the probability of committing this type of error. In this manner, the process is similar to that used for testing Type I errors, which is clarified by examples from the literature. It is hoped that researchers will use the information presented here as a means of attaching levels of probability to acceptance of null hypotheses. Am J Phys Anthropol 111:283–289, 2000.

With the advent of computer statistical packages, many anthropological studies have focused on calculating probabilities of finding differences or correlations between various biological parameters of samples drawn from human and nonhuman populations (e.g., Brothwell, 1963; Giles and Friedlaender, 1976; Cohen and Armelagos, 1984; Merbs and Miller, 1985; Gill and Rhine, 1990; Owsley and Jantz, 1994). While these probabilities (α) of committing Type I errors (i.e., rejecting null hypotheses that should be accepted) are published regularly, the converse, i.e., probabilities (β) of committing Type II errors, has not received much attention (Hodges and Schell, 1988). Type II errors are those that involve accepting null hypotheses that are false, i.e., given an alternative hypothesis, what is the probability that it (not the null hypothesis) is true? When this probability is subtracted from 1 (i.e., $1 - \beta$), statistical power is obtained; this is the probability that a test would detect differences of the magnitude set by the alternative hypothesis if those differences exist in the population from which a sample is derived.

A real-world example can clarify these concepts. While studying the relationship between stature and stress markers in prehistoric skeletal samples from the American southwest, Byers (1992) found that the "intuitively obvious" did not hold. That is, instead of persons with more stress (as exhibited by the appearance of stress markers) being shorter in stature than those without stress, there appeared to be no relationship between these two factors. For example, the analysis of the effect of the presence/absence of porotic hyperostosis on stature yielded $\alpha = 0.155$, indicating that the null hypothesis should be retained (i.e., those individuals with, and those without, porotic hyperostosis are not different in stature). This then leads to the question of how big a difference would be expected if there was a relationship between these two factors. Howe and Schiller (1952) showed a 2% decrease in stature of subadults in their late teens from Stuttgart who were stressed by World War II, while Malcolm (1979) showed that Papuans who were stressed during growth

*Correspondence to: Steven N. Byers, Department of Anthropology, University of New Mexico, Albuquerque, NM 87131. E-mail: j708@unm.edu

Received 2 April 1999; accepted 15 September 1999.

were 10% shorter than those who were not stressed. Thus, when 2% was used to test the Type II error rate, the result was $\beta \ll 0.05$, which indicates a low probability that the skeletal samples were drawn from a population of stressed and unstressed people whose statures differed by 2%. Inversely, there is a very high probability (power $\gg 0.95$) that this test would detect differences of 2% between stressed and unstressed peoples. Arguably, this finding is more interesting than the reverse, because it is counterintuitive. However, negative results such as these often go unreported because of the difficulty in assigning a probability level to the statement that there are no differences. Therefore, this paper will demonstrate how Type II error testing can be implemented in a manner similar to the process for testing Type I errors in biological anthropology studies. Thus readers, finding themselves on familiar ground, may be more likely to concentrate on β, especially in post-test situations.

Neglect in reporting Type II errors and power in anthropological research probably is due to (at least) four problems. First, most books on statistical methods, particularly introductory texts, do not explore the concept fully (e.g., Larson, 1982; Anderson and Sclove, 1986; Snedecor and Cochran, 1989; Spatz and Johnston, 1989; Sokal and Rohlf, 1995; Madrigal, 1998). Where literally hundreds of pages describe methods for calculating probabilities of committing Type I errors using various statistical models, usually power and Type II errors are discussed in few (if any) pages. When addressed at all, they are usually discussed only at the most basic level, leaving readers to discover on their own how to apply the concept to more complex models (e.g., analysis of variance, regression). Second, in the past, tedious calculations usually were required to compute these probabilities; even now they are not part of the process for calculating Type I error rates in statistical packages, and thus are not readily available. To add to this, various workers calculate power and therefore β using differently defined parameters, thereby adding another level of complication (e.g., see Pearson and Hartley, 1951 vs. Searle, 1971 for the noncentral F-distribution). Third, many workers recommend computing power curves for various statistical tests, or power at certain arbitrary levels of differences such as "small," "medium," and "large" (Cohen, 1977; Kraemer and Thiemann, 1987; Hodges and Schell, 1988). This has two ramifications: first, the computation of Type II error probabilities is not discussed with the same intensity as power. Second, and more importantly, it creates confusion as to how results should be interpreted because researchers are confronted with multiple power values that allow for multiple Type II error probabilities. This is in marked contrast to Type I error tests, where researchers can state the probability that a sample is drawn from a population where the null hypothesis holds. Unfortunately, when using power curves, researchers cannot develop a similar single statement concerning the results of their studies when the null hypothesis is retained.

The fourth and final problem deals with bias against "negative" findings (i.e., not finding expected differences). The probability of committing a Type II error (and concomitantly power) is only important when negative results (i.e., nonsignificant differences/correlations) are encountered. Although researchers usually do not hesitate to report positive results (where power and Type II errors are irrelevant), negative results are not held in the same esteem, making it less likely that these findings will be published. As pointed out by Madrigal (1998), this underreporting of negative results has the effect of biasing our view of the world.

METHODS FOR DEALING WITH TYPE II ERRORS

Many of the above problems can be obviated by examining them more closely. Although most do not cover the concept fully, fairly complete discussions of power and Type II error testing can be found in some introductory texts (e.g., Dixon and Massey, 1969; Hays, 1973), while advanced texts do devote more time to these concepts (e.g.,

Lindman, 1992; Morrison, 1983; Neter and Kutner, 1996). Also, some books are devoted entirely to power calculation and interpretation (e.g., Cohen, 1977; Kraemer and Thiemann, 1987). However, despite this coverage, none of these texts discuss Type II error calculations with the same vigor as power, thereby leaving readers to extrapolate the concept to their own research.

The second problem of tedious calculation of β can be avoided by using functions in statistical packages, such as SAS, that calculate Type II error probabilities. This is accomplished using noncentral versions of the commonly known t, F, and χ^2 distributions (e.g., SAS uses the PROBT, PROBF, and PROBCHI functions to perform the computations for these three distributions, respectively). Only three parameters are required as input to these functions: the value from the central distribution for the probability of committing a Type I error (usually $\alpha = 0.05$), the degrees of freedom, and the noncentrality parameter. The first two values are familiar to all who have done statistical testing. The noncentrality parameter (variously referred to as ϕ, ϕ^2, λ, and δ) is the ratio between the expected correlation or size difference and its error. Unfortunately, it is the calculation of this parameter that differs among different workers (e.g., see Pearson and Hartley, 1951; Searle, 1971; Rao, 1976 for this parameter in the noncentral F-distribution), thereby adding to the complexity mentioned earlier. However it is calculated, estimation of the noncentrality parameter is closely tied to alternative hypotheses, the formulation of which is the key to good Type II error testing.

The third problem of multiple power values can be obviated by calculating Type II error probabilities. In this method, instead of calculating power over a range of possible values, researchers would choose a single alternative hypothesis and test its probability of being true. This method takes advantage of the strong parallels that exist between Type I and Type II error testing, as pointed out by O'Brien (1986). Thus, researchers would formulate a null hypothesis and a specific alternative hypothesis and then calculate first the probability of committing a Type I error. If that probability is high (usually $\alpha > 0.05$), then the probability of committing a Type II error is calculated. Confronted by only one value (β), researchers are in a better position to discuss the relevance of their findings than when confronted with power curves or multiple power values at arbitrary points.

The last problem of negative results not being held in the same esteem as positive results is largely due to bias in research. Part of this bias has its foundation in the lack of complete discussions in the statistical literature of power and testing Type II error rates. Consequently, it is difficult to develop good research directed toward proving similarities and noncorrelations among groups and/or variables being studied as well as interpreting negative results when they are encountered during analysis. This is unfortunate because nonsignificance and noncorrelation can be as interesting as their opposites, especially when researchers expected to find differences or correlations but did not. What most researchers face at this point is how to "prove" their conclusion (e.g., that the groups are not significantly different).

This difficulty can be overcome by choosing anthropologically feasible alternative hypotheses and testing for the probability of committing Type II errors. If these probabilities are low, there is statistical reason to believe that the groups and/or variables being studied do not differ by values of the same (or greater) magnitude that would be expected if statistical significance had been demonstrated. Thus, the problems of "proving" nonsignificance and the confusion surrounding multiple powers of tests are eliminated. The only major obstacle is the generation of an alternative hypothesis.

In most research situations, there is an anthropological question that is being studied (e.g., two or more populations being researched are different in some way) from which is developed statistical hypotheses for testing (e.g., if the populations are different, then the means of some of their traits will be significantly different from a statistical standpoint). The statistical hypothesis usually is expressed in the format of decision

theory as in the following well known format:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_A: \mu_1 - \mu_2 \neq 0$$

Although the development of an alternative hypothesis (either anthropological or statistical) for testing nonsignificance is more complex, it follows the same pattern. That is, if the populations and/or traits under study are not statistically significant, what difference would be considered anthropologically important enough that researchers would want to be able to detect it? This would lead to a statistical alternative hypothesis that could be tested in the same manner as the null hypothesis. Again, the parallel between Type I and Type II error testing is evident. Where the researcher has an anthropological expectation that leads to the statistical null hypothesis, there is another anthropological expectation that leads to a statistical alternative hypothesis.

Information for developing an anthropological (and, consequently, statistical) alternative hypothesis originates from the findings of related research. In some cases, the study being undertaken provides an expected difference or correlation, as when males and females are analyzed separately and the null hypothesis is rejected in one but retained in the other. The significant difference or correlation in the one sex provides a logical basis for determining an alternative hypothesis for the other sex. In other cases, related but separate research results are used, as in the case when a number of groups are not found to be significantly different but previous research into similar groups was found to differ. This previous research would provide the types of deviations expected for the alternative hypothesis of the groups being studied.

EXAMPLES AND CONCLUSION

By way of examples, the above concepts will be applied to several research situations using SAS software to calculate the probability of committing a Type II error. Littleton and Frohlich (1993) present a number of *t*-tests comparing tooth wear rates among four different prehistoric peoples from the Arabian Gulf area. A comparison of the average attrition rate for molars between the Umm an-Nar and Islamic groups showed that their maxillary molars were significantly different in wear ($\alpha < 0.005$), but not their mandibular molars ($\alpha = 0.50$). Since this finding is inconsistent with what is expected, one might want to know the probability that the mandibular wear scores were drawn from a population characterized by the differences seen between the maxillary molars. Their data indicate that this difference is (3.61 − 2.46 =) 1.15; thus, a formal statement of the null and alternative hypotheses for the mandibular molar wear would be:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_A: \mu_1 - \mu_2 = 1.15$$

As mentioned above, to test the alternative hypothesis, three parameters are needed: the value from the central distribution for the probability of committing a Type I error, the degrees of freedom (df), and the noncentrality parameter. The value from the central *t*-distribution for $\alpha = 0.05$ can be obtained from a table or from a statistical package function (e.g., TINV in SAS) by knowing α and df. The degrees of freedom are dependent on the sample sizes of the two groups; in this case, n = 44 for the Umm an-Nar and n = 74 for the Islamic peoples. Finally, the noncentrality parameter (as used by SAS) can be obtained from the following formula:

$$\phi = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (1)$$

where: ϕ = noncentrality parameter
μ_1 = mean of population 1
μ_2 = mean of population 2
σ_1 = variance of population 1
σ_2 = variance of population 2
n_1 = sample size from population 1
n_2 = sample size from population 2.

Thus, using the value of the alternative hypothesis as well as the degrees of freedom and the variances of the two groups (Umm

an-Nar: $s^2 = 1.55^2$; Islamic: $s^2 = 1.63^2$), ϕ can be calculated as:

$$\phi - \frac{1.15}{\sqrt{\frac{1.55^2}{44} + \frac{1.63^2}{74}}} = 3.80. \qquad (2)$$

From this information, two SAS statements can be constructed. The first gives the value from the central *t*-distribution using the TINV function of SAS; this function requires the inverse of the Type II error rate (i.e., $1 - \alpha = 0.95$), and the df = $(44 - 1) + (74 - 1) = 116$ (this is used instead of a tabular value, for ease of computation). The second statement calculates the probability of committing a Type II error from the results of the first statement, the degrees of freedom, and the value of ϕ. The commands are:

```
TCALC = TINV(.95,116);
P = PROBT(TCALC,116,3.81);
```

Execution of these statements yields a result of $\beta = 0.016$; thus, the conclusion by Littleton and Frohlich (1993) that the groups did not differ in the amounts of their lower molar wear is supported.

Calculation of Type II error rates in the correlation/regression situation follows the same pattern. Since null hypotheses state that there is no significant correlation/regression, Type II error computations must determine the probability that the slope or correlation coefficient reaches or exceeds an expected value. For example, when studying sexual dimorphism in the pelvis, Walrath and Glantz (1996) found a significant regression between the bispinous diameter and the square of the diameter of the femoral head in human females. However, since they did not find a similar significance in males, it is reasonable to use the slope of the female regression as a test of the probability that the male values came from a population characterized by the same slope.

In the case of regression, the noncentrality parameter can be calculated from the following equation:

$$\phi = \frac{b_1}{s_{b_1}} \qquad (3)$$

ϕ = noncentrality parameter
b_1 = expected slope
s_{b_1} = standard error of the slope.

As can be seen, two values are needed: the expected slope (i.e., that of the females) and the standard error of the slope (i.e., that of the male slope). From their data, the slope of the females is approximately 0.019, while the standard error of the male slope is 0.0036. Thus, the noncentrality parameter can be calculated using formula (3) as:

$$\phi = \frac{.019}{.0036} = 5.28. \qquad (4)$$

Since n = 40 males, the SAS statements for calculating the Type II error rate are:

```
TCALC = TINV(.95,39);
P = PROBT(TCALC,39,5.28);
```

This yields $\beta = 0.0004$; thus, there is a low probability that the males came from a population of values whose slope is of the same magnitude as that seen in females, and the authors' conclusion that males differ from females in this respect is bolstered.

In analysis of variance (ANOVA), the situation is more complex. Since the null hypothesis in this model is that the means of all populations studied are equal, the amount by which they are unequal must be determined to construct a reasonable alternative hypothesis. Although not as intuitive as the previous examples, the situation is still the same. For example, Byers (1992) hypothesized that, if stress occurred late in an individual's development, statural decrease would be less likely to be corrected before maturation and cessation of growth. He used the presence of enamel hypoplasias to demonstrate stress during growth, and tested the hypothesis that late stress would cause a greater decrease, using an ANOVA of three groups: no hypoplasias (nh), hypoplasias occurring during ages 1–10 (h1–10), and hypoplasias occurring at 11 years and older (h11+) against stature. Since he discovered that there was no difference between the three groups ($\alpha = 0.207$), he wished to determine the probability that the statures from his sample were drawn from a population of individuals characterized by the dif-

ference in height seen between stressed and unstressed extant populations (as mentioned above, at least 2% and as high as 10%).

The noncentrality parameter for the noncentral F-distribution as used by SAS is:

$$\phi = \frac{\sum_{i=1}^{g} n_i(\mu_i - \mu.)^2}{\sigma^2} \quad (5)$$

where: ϕ = noncentrality parameter
n_i = sample size of group i from g groups
μ_i = mean of group i
$\mu.$ = average of group means
σ^2 = variance within groups.

As can be seen, this equation requires the group sample sizes, the expected differences between the group and overall means, and the within-groups variance (estimated by s^2_w). The overall mean stature of the sample is 1,610 mm; thus, it is expected that the stature of stressed individuals is no greater than (1,610 * 0.98 =) 1,578 mm and could be as short as (1,610 * 0.9 =) 1,490 mm. Additional values are: s^2_w = 2,772.14, n_{nh} = 37; n_{h1-10} = 82; and n_{h11+} = 5. Since there is a range of possible statural decreases but there must be specific values for the three groups, the values chosen correspond to a 2% and 4% decrease for the n_{h1-10} and n_{h11+} groups, respectively. The reasoning for these choices is that a 2% decrease is minimum and therefore expected of those stressed earlier in life, and those stressed later should have at least twice as much of a decrease. This leads to statural values of (1,610 * 0.98 =)(1,578 for the n_{h1-10} group and a stature of (1,610 * 0.96 =) 1,546 for the n_{h11+} group. This leads to an overall mean of: (1,610 + 1,578 + 1,546)/3 = 1,578.

Using the above values, the noncentrality parameter is:

$$\phi = \frac{37(1610 - 1578)^2 + 82(1578 - 1578)^2 + 5(1546 - 1578)^2}{2772.14} = 15.51. \quad (6)$$

This leads to the SAS commands:

```
FCALC = FINV(.95,2,120);
P = PROBF(FCALC,2,120,15.51);
```

Execution of these statements gives a value of $\beta = 0.054$, which indicates that the there is a low probability that the sample was drawn from a population characterized by similar levels of stress seen in extant populations.

In sum, testing Type II error rates can be made to mirror the process for testing Type I errors. By the careful selection of an anthropologically meaningful alternative hypothesis, researchers can test whether the samples being studied would deviate by the amount expected if the differences or correlations being studied were present. In this manner, although researchers cannot state that the parameters studied are not different, they can apply a probability to the statement that the samples did not deviate by an anthropologically significant difference. Since the results of tests with $\beta < 0.05$ and $\alpha > 0.05$ are important and should be published, it is hoped that this information will aid researchers in attaching a level of probability when accepting null hypotheses.

LITERATURE CITED

Anderson TW, Sclove SL. 1986. The statistical analysis of data, 2nd ed. Palo Alto, CA: Scientific Press.

Brothwell DR. 1963. Dental anthropology. Oxford: Pergamon Press.

Byers SN. 1992. The relationship between stress markers and adult body size. Unpublished Ph.D. dissertation. University of New Mexico, Albuquerque, NM.

Cohen J. 1977. Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cohen MN, Armelagos GJ. 1984. Paleopathology at the origins of agriculture. Orlando, FL: Academic Press.

Dixon WJ, Massey FJ. 1969. Introduction to statistical analysis. New York: McGraw-Hill, Inc.

Giles E, Friedlaender JS. 1976. The measure of man: methodologies in biological anthropology. Cambridge, MA: Peabody Museum Press.

Gill GW, Rhine RS. 1990. Skeletal attribution of race: methods for forensic anthropology. Anthropological papers, no. 4. Albuquerque, NM: Maxwell Museum of Anthropology, University of New Mexico.

Hays W. 1973. Statistics for the social sciences, 2nd ed. New York: Holt, Rinehart and Winston, Inc.

Hodges DC, Schell LM. 1988. Power analysis in biological anthropology. Am J Phys Anthropol 77:175–181.

Howe PE, Schiller M. 1952. Growth responses of school children to changes in diet and environmental factors. J Appl Physiol 5:51–61.

Kraemer HC, Thiemann S. 1987. How many subjects? Newbury Park, CA: Sage Publications.

Larson HJ. 1982. Introduction to probability theory and statistical inference. New York: John Wiley and Sons.

Lindman HR. 1992. Analysis of variance in experimental design. New York: Springer-Verlag.

Littleton J, Frohlich B. 1993. Fish-eaters and farmers: dental pathology in the Arabian Gulf. Am J Phys Anthropol 92:427–447.

Madrigal L. 1998. Statistics for anthropology. Cambridge, UK: Cambridge University Press.
Malcolm L. 1979. Protein-energy malnutrition and growth. In: Faulkner T, Tanner JM, editors. Human growth 3: neurobiology and nutrition. New York: Plenum Press. p 361-372.
Merbs CF, Miller RJ. 1985. Health and disease in the prehistoric southwest. Anthropological research papers, no. 34. Tempe, AZ: Arizona State University.
Morrison DF. 1983. Applied linear statistical methods. Englewood Cliffs, NJ: Prentice-Hall, Inc.
Neter J, Kutner MH. 1996. Applied linear statistical models, 4th ed. Homewood, IL: Richard D. Irwin, Inc.
O'Brien RG. 1986. Power analysis for linear models. In: Proceedings of the Eleventh Annual SAS Users Group International Conference. p 915–922.
Owsley DW, Jantz RL. 1994. Skeletal biology in the great plains: migration, warfare, health, and subsistence. Washington, DC: Smithsonian Institution Press.
Pearson ES, Hartley HO. 1951. Charts of the power function for analysis of variance tests, derived from the non-central F-distribution. Biometrika 38:112–130.
Rao CR. 1976. Linear statistical inference and its applications. New York: John Wiley and Sons.
Searle SR. 1971. Linear models. New York: John Wiley and Sons.
Snedecor GW, Cochran WG. 1989. Statistical methods, 8th ed. Ames, IA: Iowa University Press.
Sokal RR, Rohlf FJ. 1995. Biometry, 3rd ed. New York: W.H. Freeman and Co.
Spatz C, Johnston JO. 1989. Basic statistics, 4th ed. Pacific Grove, CA: Brooks/Cole Publishing Co.
Walrath DE, Glantz MM. 1996. Sexual dimorphism in the pelvic midplane and its relationship to Neanderthal reproductive patterns. Am J Phys Anthropol 100:89–100.